AD_____

Award Number: W81XWH-10-1-0589

TITLE: Global Characterization of Protein Altering Mutations in Prostate Cancer

PRINCIPAL INVESTIGATOR: Jay Shendure, M.D., Ph.D.

CONTRACTING ORGANIZATION: University of Washington
Seattle, WA 98195

REPORT DATE: August 2011

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
August 2011	Annual	15 July 2010 – 14 July 2011

4. TITLE AND SUBTITLE

Global Characterization of Protein Altering Mutations in Prostate Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0589

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jay Shendure, M.D., Ph.D.

E-Mail: shendure@uw.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Washington
Seattle, WA 98195

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The research supported by this award aims to use a new generation of technologies for DNA sequencing to comprehensively scan the genomes of a series of prostate cancers for small mutations that disrupt protein-coding sequences. Our specific aims are as follows: (1) To carry out the genome-wide identification of nonsynonymous mutations in a limited number of prostate metastases using second-generation technologies for targeted capture and sequencing; (2) To evaluate the mutational histories of individual mutations within the progression of the cancer in which it was observed, and to assess the prevalence of candidate cancer genes observed here in prostate cancer. (3) To perform integrative analyses of somatic mutation with gene expression and copy number change data collected on the same samples. To date, we have performed high-quality whole exome sequencing of 23 prostate cancers derived from 16 different lethal metastatic tumors and 3 high grade primary carcinomas. We have found that a subset of prostate cancers that exhibit a clear "hypermutator" phenotype with respect to point mutations, with potential implications for resistance to cancer therapeutics. We have also identified a subset of genes that appear to be recurrently mutated in our discovery set, and are therefore strong candidates for additional investigation for potential relevance to prostate cancer initiation or progression to metastasis.

15. SUBJECT TERMS

prostate cancer, genome, exome, sequencing, mutation,

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	11	USAMRMC

Table of Contents

Introduction

The identification of recurrent, protein-altering genetic alterations is frequently the means by which a given gene is initially implicated in tumor biology. However, we currently lack anything approaching a comprehensive picture of the protein-altering mutations that are biologically relevant or potentially specific to prostate cancer. The research supported by this award aims to use a new generation of technologies for DNA sequencing (Shendure and Ji 2008) to comprehensively scan the genomes of a series of prostate cancers for small mutations that disrupt protein-coding sequences. Our specific aims are as follows: (1) To carry out the genome-wide identification of nonsynonymous mutations in a limited number of prostate metastases using second-generation technologies for targeted capture and sequencing; (2) To evaluate the mutational histories of individual mutations within the progression of the cancer in which it was observed, and to assess the prevalence of candidate cancer genes observed here in prostate cancer. (3) To perform integrative analyses of somatic mutation with gene expression and copy number change data collected on the same samples.

Body

This is a "synergy" project between the laboratories of Dr. Jay Shendure in the Department of Genome Sciences at the University of Washington (UW) and Dr. Peter Nelson in the Division of Human Biology at the Fred Hutchinson Cancer Research Center (FHCRC). Because these are separate awards to the two investigators, this progress report is specific to tasks from the statement of work (SOW) assigned to the Shendure Lab only (or to progress within the Shendure Lab for joint tasks). As per the instructions, progress is reported in association with each task listed in the SOW. Only tasks containing a UW component are listed here.

Aim 1: Perform a comprehensive screen for protein coding alterations in prostate metastases.

Task 3. DNA isolation and shotgun library construction (Months 1-10) [UW]

In Year 1 of this project, we performed whole exome sequencing of 23 prostate cancers derived from 16 different lethal metastatic tumors and 3 high grade primary carcinomas using solution-based hybrid capture (Nimblegen) followed by massively parallel sequencing (Illumina). All tumors were propagated in mice as xenografts, designated the LuCaP series. Genomic DNA was isolated from frozen tissue blocks using the QIAGEN DNeasy Blood and Tissue kit. Shotgun libraries were constructed by shearing gDNA, ligating sequencing adaptors, and performing PCR amplification.

Task 4. Array-based enrichment of coding sequences (Months 7-13) [UW]

The Nimblegen EZ SeqCap kit (Roche) was used as recently described (O'Roak, Deriziotis et al. 2011) in order to capture subsequences of the genome corresponding to coding regions, *i.e.* the "exome". Shotgun libraries were hybridized to either the EZ SeqCap V1 or V2 solution-based probes, and amplified. V1 probes (used in eight samples) targeted 26.6 Mb corresponding to the CCDS definitions of exons, while V2 probes (used in 15 samples) targeted 36.6 Mb corresponding to the RefSeq gene database.

Task 5. Massively parallel sequencing of tumor and control exomes (Months 10-16) [UW]

Post-enrichment libraries for these 23 prostate cancers were sequenced on either the Illumina GAIIx or HiSeq platforms (**Table 1**).

Table 1: Methods used to capture and sequence prostate cancer exomes. We used two versions of Nimblegen EZ SeqCap capture probes in this study. Eight samples were captured using V1 probes (targeting the 26.6 Mb Consensus Coding Sequence Database (CCDS), while the remainder of samples were captured using V2 probes (targeting the 36.6 RefSeq database). Four samples were indexed with barcodes prior to capture and sequencing. V1, Nimblegen V1 solution capture probes targeting CCDS coordinates; V2, Nimblegen V2 solution capture probes targeting RefSeq coordinates ; PE-76, paired-end sequencing using 76 bp reads; PE-100 paired-end sequencing using 100 bp reads.

Sample ID	Capture Method	Indexing	Sequencer	Run-type
LuCaP 23.1	V2	no	HiSeq	PE-100
LuCaP 23.12	V1	no	Illumina GAIIx	PE-76
LuCaP 23.1AI	V1	no	Illumina GAIIx	PE-76
LuCaP 35	V1	no	Illumina GAIIx	PE-76
LuCaP 35V	V1	no	Illumina GAIIx	PE-76
LuCaP 49	V1	no	HiSeq	PE-100
LuCaP 58	V2	no	HiSeq	PE-100
LuCaP 70	V2	no	HiSeq	PE-100
LuCaP 73	V2	yes	HiSeq	PE-100
LuCaP 77	V2	yes	HiSeq	PE-100
LuCaP 78	V2	no	HiSeq	PE-100
LuCaP 81	V2	no	HiSeq	PE-100
LuCaP 86.2	V1	no	HiSeq	PE-100
LuCaP 92	V2	no	HiSeq	PE-100
LuCaP 93	V2	no	HiSeq	PE-100
LuCaP 96	V1	no	Illumina GAIIx	PE-76
LuCaP 96AI	V1	no	Illumina GAIIx	PE-76
LuCaP 105	V2	no	HiSeq	PE-100
LuCaP 115	V2	no	HiSeq	PE-100
LuCaP 136	V2	no	HiSeq	PE-100
LuCaP 141	V2	no	HiSeq	PE-100
LuCaP 145.2	V2	yes	HiSeq	PE-100
LuCaP 147	V2	no	HiSeq	PE-100

Task 6. Read mapping, variant calling, and mutation annotation (Months 11-17) [UW]

We dealt with the possibility of mouse genomic DNA contamination by mapping sequence reads to both the human (UCSC hg18) and mouse (mm9) genome sequences using BWA (Li and Durbin 2009). Reads that mapped to the mouse genome were excluded from further analysis. See **Figures 1**, **2**, and **3**

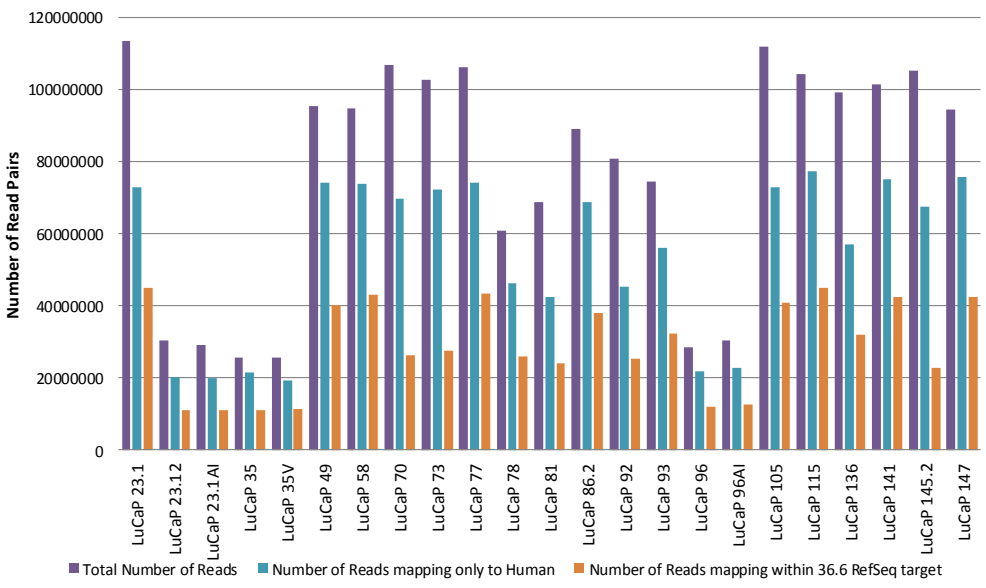

Figure 1: Summary of mapping statistics across 23 PCa xenograft exomes. LuCaP samples 96, 96AI, 23.12, 23.1AI, 35 and 35V were sequenced using the Illumina GAIIx, which accounts for the smaller number of reads obtained for these samples

for mapping statistics and calculations of mapping complexity. Sequence variant calls were performed by *samtools* (Li, Handsaker et al. 2009) after removing potential PCR duplicates, and were filtered to consider only positions with more than 8x coverage and a Phred-like consensus quality of 30 (Ng, Turner et al. 2009). To eliminate common germline polymorphisms from consideration, variants that had the same position as variants present in pilot data from the 1000 Genomes Project or in ~2,000 exomes corresponding to normal (non-tumor, non-xenografted) tissues sequenced at the University of Washington were removed from consideration. Genotypes were annotated using the SeattleSeq server (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) and only nonsynonymous variants (missense/nonsense/splice-site mutations) were considered in identifying genes with recurrent mutations. The subset of genes that were recurrently mutated was then validated manually using IGV, the Integrated Genomics Viewer, to identify and remove false positive calls due to the presence of an insertion/deletion or incorrectly mapping read (12).

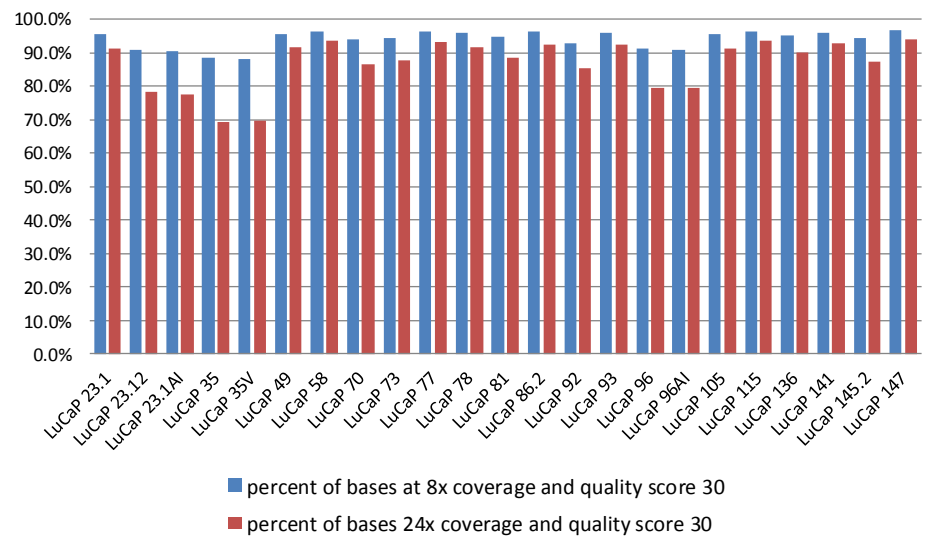

Figure 2: Fraction of bases in the V1 target definition that were covered to sufficient depth to enable basecalling.

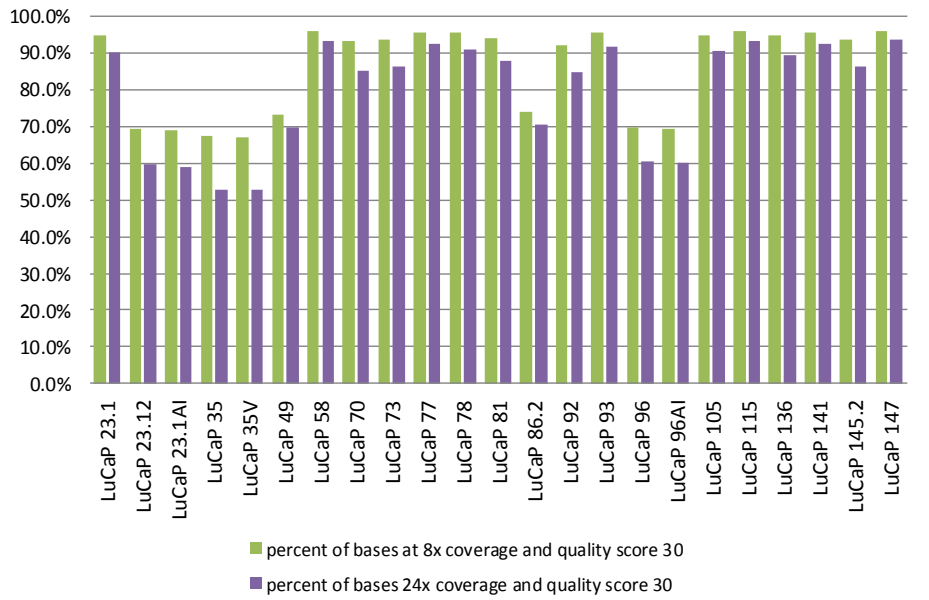

Figure 3: Fraction of bases in the V2 target definition that were covered to sufficient depth to enable basecalling. Samples LuCaP 96, 96V, 23.12, 23.1AI, 35, 35V, 49 and 86.2 were selected for a smaller (V1) target, which accounts for their relatively lower coverage of these regions.

Aim 2: Evaluate mutational histories and prevalence screen of candidate cancer genes.

Task 10. Application to evaluate mutation histories (Months 20-24) [UW]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Task 11. Application to prevalence screen of candidate cancer genes (Months 18-30) [UW]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Task 12. Read mapping, variant calling, and mutation annotation (Months 21-31) [UW]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Task 13. Verification/confirmation of sequence variants (Months 20-26) [UW]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Aim 3: Integrate analyses of molecular alterations in metastatic and primary prostate cancer.

Task 15. Analysis of mutational patterns in metastatic prostate cancer (Months 11-17) [UW]

To date, our analyses of mutational patterns have produced two key results. First, we have found that a subset of prostate cancers exhibit a "hypermutator" phenotype with respect to point substitution mutations. Second, we have identified a set of candidate genes that may be recurrently mutated in prostate cancer. These analyses are discussed separately below.

Prostate cancers with "hypermutated" genomes. We observed that the exomes of three prostate cancers, LuCaP 58, LuCaP 73 and LuCaP 147 possessed a strikingly high number of novel, non-synonymous single nucleotide variants, nearly tenfold more than other tumors (p=0.0097) (**Figure 4**). There were no distinctive features to suggest why these tumors should have more variants. Each tumor originated as a high grade Gleason 9 cancer, all were from individuals of Caucasian ancestry, one represented a primary neoplasm, one a lymph node metastasis, and one a metastasis to the liver. We hypothesized that the large number of nov-SNVs observed in three prostate cancers may be due to a „mutator phenotype" that developed during the initial stages of tumorigenesis, as a consequence of therapeutic pressures and subsequent clonal selection, or evolved while passaged in the mouse hosts. To determine if these results reflect truly elevated numbers of somatic mutations within human tumors and not as a result of passage within mice, we sequenced paired normal and directly resected, non-xenografted, tumor samples corresponding to one hypermutated xenograft (LuCaP 147), and two non-hypermutated xenograft lines (LuCaP 92 and LuCaP 145.2). Of 2,368 novSNVs in LuCaP147 able to be called across all three samples (xenograft, derivative tumor and normal tissue) 1,402 were somatic and present with metastasis tissue. In contrast, the other two non-xenografted tumors (corresponding to LuCaP 92 and LuCaP 145.2) had between 31 and 58 somatic mutations. Furthermore, because we sequenced a neighboring metastasis, rather than the exact metastasis from which LuCaP147 was derived, the result indicates that at least these ~1,400 somatic mutations were shared between these metastases. The vast majority of the ~1,300 novSNVs observed in the LuCaP147 xenograft but not the metastasis likely occurred during passaging within mice, or were specific to the metastasis from which LuCaP147 was derived.

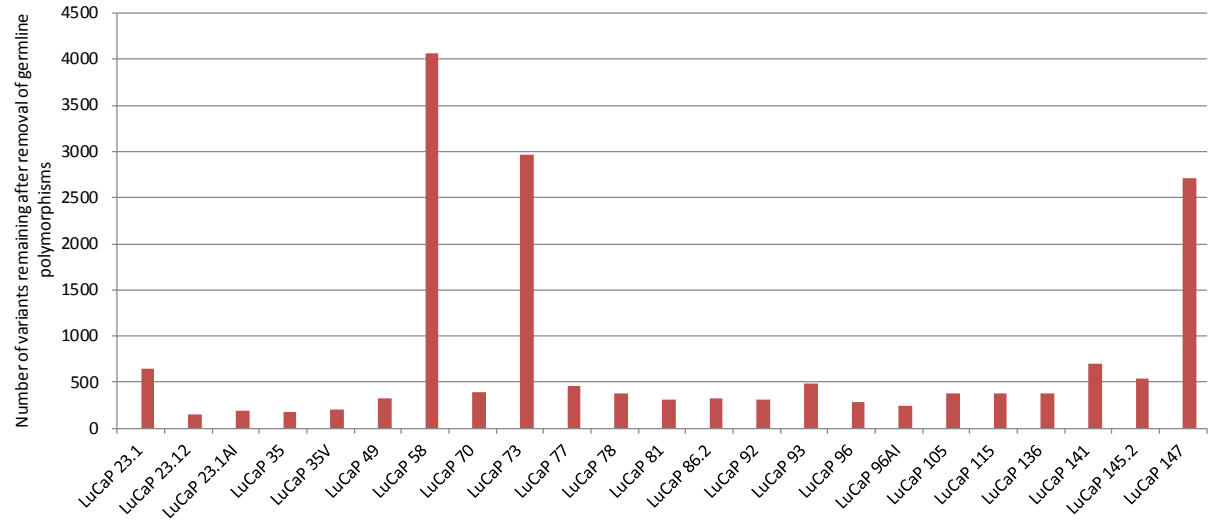

Fig. 4. A subset of xenografts exhibit a high number of mutations. After filtering to remove common germline polymorphisms, three xenografts (LuCaP 73, LuCaP 147 and LuCaP 58) exhibit a "hypermutated" phenotype, with several thousand novel SNVs each. This contrasts with the other 20 xenografts, which have 362 +/- 147 coding alterations remaining after filtering.

Recurrent nonsynonymous genomic sequence alterations in prostate cancers. We examined the set of novel, nonsynonymous single nucleotide variants (nov-nsSNVs) to identify those genes that may be recurrently affected by protein-altering point mutations across different tumors. In order to reduce spurious findings due to inconsequential passenger mutations, we excluded the three "hypermutated" tumors from this analysis. We also manually examined read pileups for variants in genes with potential recurrence attributable to basecalling artifacts due to either insertions/deletions or poorly mapping reads. Across 16 tumors from unrelated individuals, 131 genes had nov-nsSNVs in two or more exomes, and 23 genes had nov-nsSNVs in three or more exomes. A subset of the novel variants are likely due to instances where very rare germline variants (i.e. not seen in several thousand other chromosomes) occur in the same gene, as we cannot distinguish these from somatic mutations. We therefore excluded from consideration the 1% of genes with the highest rate of very rare germline variants, i.e. singletons, based on an analysis of control exomes (as some genes are much more likely to contain very rare germline variants than other genes) (Bustamante, Fledel-Alon et al. 2005; Lohmueller, Indap et al. 2008). This reduced the number of candidates to 104 genes with nov-nsSNVs in two or more exomes, and 12 genes with nov-nsSNVs in three or more exomes. To further segregate candidate genes with the goal of identifying those with recurrent somatic mutations, we estimated the probability of recurrently observing germline nov-nsSNVs in each candidate gene by iterative sampling from 1,865 other exomes sequenced at the University of Washington. We excluded from consideration genes for which the probability of observing the genes recurrently mutated due to germline variation was greater than 0.001. This reduced the number of candidates to 20 genes with nov-nsSNVs in two or more exomes, and 10 genes with nov-nsSNVs in three or more exomes (**Table 2**). Notably, whereas we began with 4 genes with nov-nsSNVs in four or more exomes (MUC16, SYNE1, UBR4, and TP53), only one of these (TP53) remained in our final candidate list, where it is the most significant. These data and analysis provide a strong set of candidates for further investigation.

# of samples	Gene ID	Gene Name	P-value of being germline	Individual mutations seen
5	TP53	tumor protein p53 (Li-Fraumeni syndrome)	< 0.00005	LuCaP73(ARG306GLN), LuCaP136(ARG280stop), LuCaP96AI(CYS238TYR), †LuCaP92(GLU198stop), LuCaP73(ARG175CYS), LuCaP70(TYR163HIS), LuCaP77(PRO278SER)
3	SDF4	stromal cell derived factor 4	< 0.00005	LuCaP108(ASP276ASN), LuCaP78(GLY76SER), LuCaP115(ALA9SER)
3	PDZRN3	PDZ domain containing RING finger 3	< 0.00005	LuCaP96AI(ARG727CYS), LuCaP108(GLY570SER), LuCaP73(ARG463CYS), LuCaP92(ILE331LEU)
3	DLK2	delta-like 2 homolog	0.00005	LuCaP70(ARG371HIS), *LuCaP145.2(SER361ARG), LuCaP23.1AI(HIS280GLN)
3	FSIP2	fibrous sheath interacting protein 2	0.00005	LuCaP81(LYS22ASN), †LuCaP92(THR698ILE), LuCaP136(GLN1526HIS)
3	NRCAM	neuronal cell adhesion molecule	0.00015	LuCaP115(MET1094ILE), LuCaP86.2(LYS645GLU), †LuCaP145.2(SER329CYS)
3	MGAT4B	mannosyl GAT4B	0.0002	LuCaP108(ALA504THR), LuCaP96AI(ARG168CYS), LuCaP136(VAL150MET)
3	PCDH11X	protocadherin 11 X-linked	0.0003	*LuCaP145.2(VAL38PHE), LuCaP58(MET867VAL), LuCaP108(VAL1007ILE), LuCaP49(THR1296ASN)
3	GLI1	glioma-associated oncogene homolog 1 (zinc finger protein)	0.0003	LuCaP86.2(ARG20TRP), LuCaP78(ARG81GLN), LuCaP23.1AI(PRO210THR)
3	KDM4B	Lysine-specific demethylase 4B	0.00035	LuCaP73(ALA265VAL), LuCaP108(ARG534TRP), LuCaP35V(ALA555VAL), LuCaP73(ALA827VAL), LuCaP86.2(SER1036CYS)
2	DKK1	dickkopf homolog 1	< 0.00005	†LuCaP92(GLU151GLN), LuCaP93(SER244TYR)
2	RAB32	RAB32, member RAS oncogene family	0.00005	LuCaP93(VAL66ILE), LuCaP141(SER109stop)
2	PLA2G16	phospholipase A2, group XVI	0.00015	LuCaP115(SER85LEU), LuCaP35V(PRO19HIS)
2	TFG	TRK-fused gene	0.00015	LuCaP96AI(ASN134HIS), LuCaP141(GLN318stop), LuCaP147(TYR319stop)
2	TBX20	T-box 20	0.0002	LuCaP77(ARG437HIS), LuCaP23.1AI(ALA52SER)
2	ZNF473	zinc finger protein 473	0.00025	LuCaP108(VAL465ILE), LuCaP115(GLY652ARG)
2	SF3A1	splicing factor 3a, subunit 1, 120kDa	0.0006	LuCaP70(PRO558LEU), LuCaP23.1AI(VAL479ILE)
2	NMI	N-myc (and STAT) interactor	0.00075	LuCaP141(ILE302ARG), LuCaP86.2(GLN101ARG)
2	IKZF4	IKAROS family zinc finger 4 (Eos)	0.0008	LuCaP93(ASP106ASN), LuCaP81(ASP498ASN)
2	BDH1	3-hydroxybutyrate dehydrogenase, type 1	0.00095	LuCaP73(VAL190ILE), LuCaP96AI(THR176MET), LuCaP147(VAL142ILE), LuCaP115(HIS74TYR), LuCaP147(ALA50VAL)

Table 2: Genes with recurrent novel, nonsynonymous alterations. P-values were estimated by randomly sampling from 1,865 other exomes sequenced at the University of Washington to estimate the probability of recurrently observing nov-nsSNVs in a given candidate gene. These are the 20 genes with the best p-values. *The nov-nsSNV in this gene was determined to be a rare germline mutation within this xenograft. †The nov-nsSNV in this gene was determined to be a somatic within this xenograft.

Task 16. Analysis of mutational patterns in primary prostate cancer (Months 21-31) [UW]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Task 21. Annotating and posting project data to website (Months 18-36) [UW & FHCRC]

This task is scheduled to begin in Year 2, and has not yet been initiated.

Task 22. Completing project reports and manuscripts (Months 11-36) [UW & FHCRC]

In June 2011, we submitted a manuscript entitled "Exome Sequencing Identifies a Spectrum of Mutation Frequencies in Advanced and Lethal Prostate Cancers" to the Proceedings of the National Academy of Sciences (PNAS), describing our progress on this project to date. The reviews were generally positive, and we are currently revising the manuscript to address the comments of the reviewers. The manuscript is not included here because the paper has not yet been accepted.

Key Research Accomplishments

- We have performed high-quality whole exome sequencing of 23 prostate cancers derived from 16 different lethal metastatic tumors and 3 high grade primary carcinomas.
- We have found that a subset of prostate cancers that exhibit a clear "hypermutator" phenotype with respect to point mutations, with potential implications for resistance to cancer therapeutics.
- We have identified a subset of genes that appear to be recurrently mutated in our discovery set, and are therefore strong candidates for additional investigation for potential relevance to prostate cancer initiation or progression to metastasis.

Reportable Outcomes

- In June 2011, we submitted a manuscript entitled "Exome Sequencing Identifies a Spectrum of Mutation Frequencies in Advanced and Lethal Prostate Cancers" to the Proceedings of the National Academy of Sciences (PNAS), describing our progress on this project to date. The reviews were generally positive, and we are currently revising the manuscript to address the comments of the reviewers. The manuscript is not included here because the paper has not yet been accepted.
- Preliminary findings for this project were reported in an invited talk (Shendure) at the 17th Annual Scientific Retreat of the Prostate Cancer Foundation, (Washington DC)

Conclusion

In summary, by sequencing the exomes of 23 tumors representing a spectrum of aggressive advanced prostate cancers, we identified a large number of previously unrecognized gene coding variants with the potential to influence tumor behavior. However, our results also indicate that with notable exceptions, very few genes are mutated in a substantial fraction of tumors.

Furthermore, while the overall mutation frequencies approximate those found in other cancers of epithelial origin, we also identified a distinct subset of tumors that exhibit a hypermutated genome. It will be important to determine the mechanism(s) responsible for the enhanced point mutation rates in these malignancies, particularly if further studies demonstrate enhanced resistance to cancer therapeutics.

References

Bustamante, C. D., A. Fledel-Alon, et al. (2005). "Natural selection on protein-coding genes in the human genome." Nature **437**(7062): 1153-1157.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics (Oxford, England) **25**(14): 1754-1760.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Lohmueller, K. E., A. R. Indap, et al. (2008). "Proportionally more deleterious genetic variation in European than in African populations." Nature **451**(7181): 994-997.

Ng, S. B., E. H. Turner, et al. (2009). "Targeted capture and massively parallel sequencing of 12 human exomes." Nature **461**(7261): 272-276.

O'Roak, B. J., P. Deriziotis, et al. (2011). "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations." Nat Genet.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.